



UNIVERSIDAD  
**COMPLUTENSE**  
MADRID

Proyecto de Innovación

Convocatoria 2016/2017

Nº de proyecto: 9

Taller de empleo de Sistemas de Recuperación de código libre en el entorno  
laboral

Responsable del proyecto: Juan-Antonio Martínez-Comeche

Facultad de Ciencias de la Documentación

Departamento de Biblioteconomía y Documentación

## 1.- Objetivos propuestos en la presentación del proyecto.

En los últimos años se percibe en el entorno laboral de los egresados en Biblioteconomía y Documentación un claro aumento de las ofertas de trabajo relacionadas con aspectos tecnológicos, destacando entre ellos los puestos de Community Manager y Gestor de contenidos web (Martínez-Comeche, J. A.; et al. (2015). La demanda de profesionales de la información y documentación en España: Análisis de las denominaciones de las ofertas de trabajo y de las tareas y competencias que involucran. Revista General de Información y Documentación, vol. 25, nº 1, pp. 175-190).

Tanto en estos puestos de trabajo de reciente acuñación como en los clásicos (Documentalista, por ejemplo), las competencias relativas a la búsqueda y recuperación de información en entornos automatizados destacan entre las más demandadas en las ofertas de trabajo analizadas.

Sin embargo, el bagaje informático de los alumnos matriculados en el Grado en Información y Documentación es escaso, lo que dificulta en buena medida la asimilación de contenidos tecnológicos en general, e informáticos en particular, por parte de los alumnos de las asignaturas relativas a la Recuperación de Informaición.

Los principales objetivos perseguidos en el proyecto tratan de paliar esta situación, pudiendo resumirse de la siguiente manera:

- Incorporar a los contenidos de la asignatura ~~B~~Búsqueda y Recuperación de información+del Grado en Información y Documentación no solo los conceptos esenciales relativos a la Recuperación de información, sino el contacto y conocimiento de los principales programas y Sistemas de Recuperación de Información de código libre empleados habitualmente en organismos y empresas de todos los sectores. El conocimiento de las características de estos Sistemas de Recuperación hará a los alumnos competentes no solo en los aspectos técnicos que involucran, sino también les hará competentes a la hora de valorar cuál de dichos Sistemas de Recuperación es el más adecuado en función de las tareas automáticas requeridas en un puesto de trabajo concreto.
- Facilitar la asimilación de dichas competencias mediante la elaboración de un taller práctico, de manera que el alumno aprehenda las características de los diversos sistemas mediante la realización de tareas habituales en el entorno profesional, deduciendo por sí mismo cuál o cuáles de dichos Sistemas de Recuperación son los más adecuados. Dicho enfoque práctico es especialmente útil para los alumnos de Documentación, dado el bajo nivel técnico del que parten los alumnos de Documentación cuando afrontan por primera vez asignaturas relativas a la Recuperación de Información.

La consecución de estos objetivos mejorará el conocimiento, por parte del alumnado, de los principales programas y Sistemas de Recuperación de Información empleados en el quehacer profesional. Además, mejorará las competencias relacionadas con la Recuperación de Información por parte del alumnado con escaso bagaje técnico.

## 2.- Objetivos alcanzados.

Se han alcanzado la totalidad de los objetivos propuestos, resumidos de la siguiente manera:

- Se han incorporado a los contenidos de la asignatura **Búsqueda y Recuperación de información** del Grado en Información y Documentación durante el primer cuatrimestre del curso 2016-2017 el análisis y manejo de tres herramientas informáticas que permiten al alumnado (con escasa base tecnológica y matemática previa) la asimilación de los fundamentos y principales técnicas de recuperación automatizada de la información a través del manejo de tres sistemas reales de recuperación de información empleados en la práctica profesional e investigadora:
  - El programa Solr, que emplea por defecto los modelos booleano y vectorial, y al que se puede incorporar el modelo probabilístico.
  - El programa Terrier, que emplea por defecto el modelo DFR (Divergence From Randomness), perteneciente a la familia de modelos probabilísticos de recuperación de información.
  - El programa Indri, que emplea por defecto una mezcla de un modelo de lenguaje estadístico y un modelo de redes de inferencia.
- Se han puesto a disposición de los alumnos de la asignatura **Búsqueda y Recuperación de Información**, durante el primer cuatrimestre del curso 2016-2017, tres talleres:
  - Taller del programa Solr
  - Taller del programa Terrier
  - Taller del programa Indri

Dichos talleres incluyen los aspectos y tareas más relevantes de la recuperación de información:

- La indexación de colecciones
  - La incorporación de listados de palabras vacías específicas.
  - La incorporación de códigos concretos de stemming.
  - La recuperación de documentos de la colección mediante la introducción de consultas.
- Durante el curso 2016-2017 se ha incorporado a la docencia presencial, y por tanto ha formado parte de la evaluación de la asignatura, el taller del programa Solr, por ser uno de los Sistemas de Recuperación de código libre más ampliamente utilizado en la práctica profesional a nivel internacional.
- Durante el curso 2016-2017 se ha incorporado a la docencia, con carácter voluntario y modalidad online, formando parte igualmente de la evaluación de la asignatura, los talleres de los programas Terrier e Indri, por ser dos de los Sistemas de Recuperación de código libre más ampliamente utilizados en la investigación sobre Recuperación de Información a nivel internacional.

### 3.- Metodología empleada en el proyecto.

El plan de trabajo ha constado de las siguientes fases, cada una de ellas con su metodología específica:

- Selección de programas y sistemas de recuperación de código libre de amplia utilización en organismos y empresas de toda índole.
- Selección de tareas relacionadas con la recuperación de información más habituales en organismos y empresas de toda índole.
- Desarrollo de los contenidos de los talleres (con su posterior evaluación)

La metodología empleada en cada fase del plan de trabajo puede resumirse de la siguiente manera:

- Selección de programas y Sistemas de Recuperación de código libre. Para ello se han cumplimentado cuestionarios por parte del PAS y del alumnado que desarrolla actividades profesionales relacionadas con la Recuperación de información. La finalidad consistía esencialmente en discernir qué programas de código libre son los más ampliamente utilizados en la práctica profesional de diversos ámbitos, frente a los programas propietarios con características semejantes.
- Selección de tareas relacionadas con la recuperación de información. Para ello se han incorporado preguntas específicas en los cuestionarios completados por el personal del PAS y por los alumnos que simultanean los estudios con el quehacer profesional. La finalidad consistía esencialmente en discernir cuáles son las tareas concretas relacionadas con la búsqueda y recuperación de información que se realizan cotidianamente en los distintos puestos de trabajo, de manera que su desarrollo se incorporase a los talleres de los distintos programas seleccionados previamente.
- Desarrollo de los contenidos de los talleres. Para ello se han realizado cuestionarios a los miembros del equipo que son profesores de asignaturas vinculadas a la Recuperación de información, tanto en Facultades de Documentación como en Facultades de Informática de España, con el objetivo de recopilar datos sobre dos aspectos fundamentales: Qué programas, entre los seleccionados previamente en el ámbito laboral, son los más convenientes desde el punto de vista docente; por otra parte, en qué medida dichos programas permiten la realización de las tareas concretas que se desarrollan en el ámbito laboral. A partir de estos datos, se procedió al desarrollo de los contenidos de los talleres de Solr, Terrier e Indri, incorporando cada uno de ellos tanto la instalación en Windows como en Linux (Mac) como el procedimiento para realizar las tareas de indexación, eliminación de palabras vacías, stemming y recuperación de información.

#### **4.- Recursos humanos**

A fin de cumplir los objetivos enunciados anteriormente, se ha considerado que los recursos humanos idóneos eran los siguientes:

- Profesorado procedente de las dos ramas del conocimiento mayoritariamente involucrados en la enseñanza de la materia ~~%Recuperación de información~~ la Informática y la Documentación. Con este fin se incorporaron al equipo 4 profesores de Documentación y 2 profesores de Informática de diversas universidades españolas. Además, todos los profesores de Documentación tienen formación en el ámbito de la informática a diverso nivel.
- Alumnos de Documentación que hubiesen cursado previamente la asignatura y que trabajasen como documentalistas en diversos sectores económicos. Con este fin se incorporaron al equipo 4 alumnas que desarrollan sus trabajos en los sectores de la propiedad intelectual de tipo patrimonial, la construcción, la Administración Pública y la salud e investigación médica.
- Personal de Administración y Servicios, con el fin de incorporar el punto de vista de los profesionales que a diario gestionan la documentación y que, por tanto, conocen bien la problemática relacionada con la búsqueda y recuperación de información de carácter administrativo en el ámbito universitario. Con este fin se incorporó al equipo un miembro del PAS de la Facultad de Ciencias de la Documentación de la UCM.
- Un Informático que resolviese tanto el proceso de instalación de los programas seleccionados en los sistemas Windows y Linux, como los diversos problemas que surgiesen relacionados con las tareas específicas relacionadas con la recuperación de información.

#### **5.- Desarrollo de las actividades.**

Las principales actividades desarrolladas a lo largo del proceso han sido las siguientes:

- Selección final de los programas Solr, Terrier e Indri por satisfacer plenamente los requisitos impuestos:
  - Programas de código libre.
  - Programas de amplia utilización en los ámbitos laboral y de investigación.

- Posibilidad de elección del modelo de recuperación de información (booleano, probabilístico y vectorial), pues entre los tres desarrollan los tres modelos clásicos principales de recuperación de información: modelo booleano, modelo probabilístico y modelo vectorial.
- Facilidad en el proceso de representación e indexación de los documentos de la colección. Todos estos parámetros suelen definirse en un archivo de configuración de fácil modificación.
- Facilidad en el proceso de inclusión de tareas como la eliminación de palabras vacías o el stemming. Estas tareas suelen definirse y configurarse en un mismo archivo.
- Desarrollo de los tres talleres de Recuperación de Información:
  - Taller de Solr
  - Taller de Terrier
  - Taller de Indri

En todos los talleres se ha seguido el mismo esquema, con el fin de incorporar todos los aspectos más habituales en el ámbito profesional:

- Instalación en Windows
- Instalación en Linux (Mac)
- Proceso de indexación
- Técnica de eliminación de palabras vacías
- Técnica de stemming
- Proceso de recuperación de información

Este mismo esquema permite al alumno comparar el modo de funcionamiento, el proceso de configuración y los resultados esperables en cada uno de los sistemas incluidos, lo que facilita la decisión de elección en el futuro en un ámbito y circunstancias laborales específicas.

- Impartición a los alumnos en clase del Taller de Solr desarrollado previamente durante el primer cuatrimestre del curso 2016-2017.

- Desarrollo con carácter voluntario, por parte de los alumnos, del Taller de Terrier y el Taller de Indri, durante el primer cuatrimestre del curso 2016-2017.
- La realización de estos talleres ha sido incorporada a la calificación final de los alumnos de tercer curso del Grado en Información y Documentación matriculados en la asignatura %Búsqueda y Recuperación de Información+ durante el curso 2016-2017.
- Se ha distribuido un cuestionario a todos los alumnos de tercer curso del Grado en Información y Documentación matriculados en la asignatura %Búsqueda y Recuperación de Información+ durante el curso 2016-2017, con el fin de evaluar el Taller de Solr (el único de realización obligatoria) desde un punto de vista docente e introducir en cursos sucesivos las mejoras necesarias. Las principales conclusiones obtenidas de la distribución del cuestionario son las siguientes:
  - El Taller de Solr es útil para comprender los procesos de análisis y recuperación de información.
  - El Taller de Solr es útil para comprender el funcionamiento de un Sistema de Recuperación de Información de uso habitual en la práctica profesional.
  - Se considera conveniente dedicar menos tiempo al proceso de instalación y centrarse más en la configuración, el funcionamiento del programa y la realización de búsquedas.
- Se han realizado entrevistas a dos alumnos, con el fin de poder discernir con mayor precisión las dificultades encontradas por los alumnos en la realización de los tres talleres, de manera que se resuelvan en posteriores versiones de los mismos. Las principales conclusiones obtenidas de las entrevistas son las siguientes:
  - Se considera necesario incluir un tutorial para la instalación de Java.
  - Se considera necesario incluir un tutorial para la activación del terminal o línea de comandos.
  - Se considera necesario que los enlaces relativos a los programas que se incluyen en los talleres tengan un carácter lo más genérico posible, de manera que permitan la consulta del taller a pesar de las actualizaciones de los programas.